Text-mining the post-Soviet web

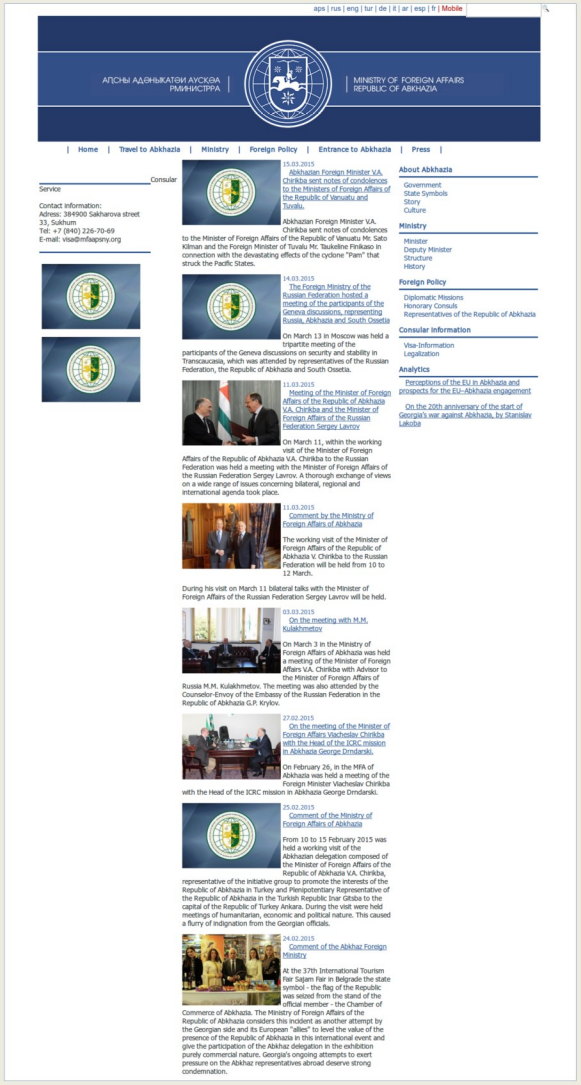
Giorgio Comai
Dublin City University
@giocomai

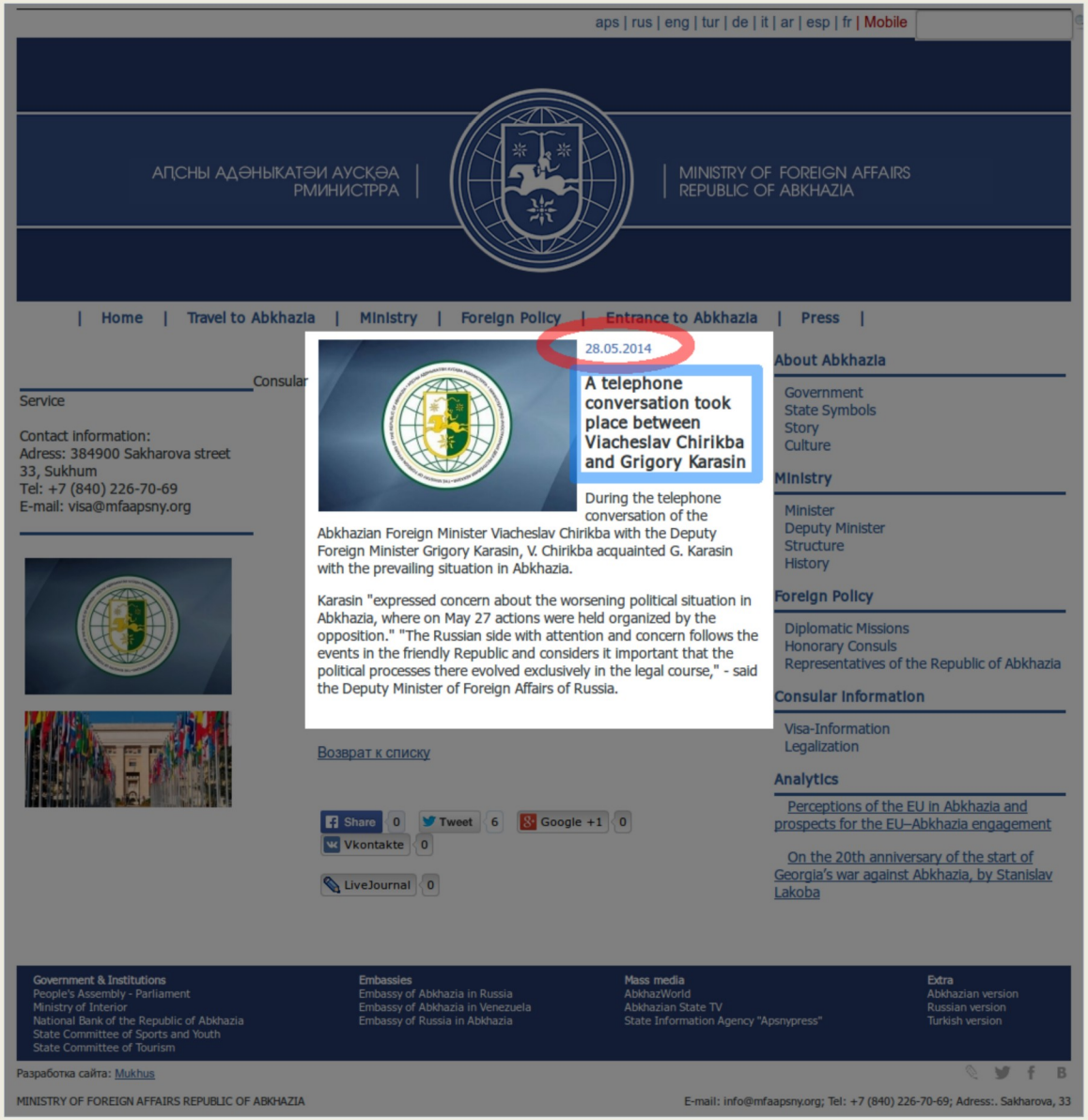
Structure of the presentation

- Creating textual datasets from the web
- Why bother?
- Some examples of results

- If it was easy enough, would area studies researchers use this approach?

Creating datasets





```
> meta(corpus[[958]])
Metadata:
  author       : Abkhazia's MFA
  datetimestamp: 2014-05-28
  description  : character(0)
  heading      : A telephone conversation took place between Viacheslav Chirikba and Grigory Karasin
  id           : ABK2461
  language     : en
  origin       : http://mfaapsny.org/en/information/?ID=2461
> corpus[[958]]
<<PlainTextDocument (metadata: 7)>>
28.05.2014
A telephone conversation took place between Viacheslav Chirikba and Grigory Karasin
During the telephone conversation of the Abkhazian Foreign Minister Viacheslav Chirikba with the Dep
uty Foreign Minister Grigory Karasin, V. Chirikba acquainted G. Karasin with the prevailing situatio
n in Abkhazia.
Karasin "expressed concern about the worsening political situation in Abkhazia, where on May 27 acti
ons were held organized by the opposition." "The Russian side with attention and concern follows the
events in the friendly Republic and considers it important that the political processes there evolve
d exclusively in the legal course," - said the Deputy Minister of Foreign Affairs of Russia.
```

Export, archive, etc.

B	C	D	E	F	G	H	I
nameOfP	nameOfW	dates	artic	titles	language	articlesLinks	articlesTxt
MFAs	Abkhazia	2012-04-16	1	On April 16-18- the Minister of Foreign Affai	en	http://mfaapsny.org/en/information/?ID=44	16.04.2012 On April 16-18, the Minister of Foreign Affairs of the Republic of Ab On April 16-18, the Minister of Foreign Affairs of the Republic of Ab During his stay in Moscow, Viacheslav Chirikba will take part in th During the visit, the Foreign Minister of Abkhazia also plans to visi
MFAs	Abkhazia	2012-04-17	2	The Minister of Foreign Affairs of the Repub	en	http://mfaapsny.org/en/information/?ID=47	17.04.2012 The Minister of Foreign Affairs of the Republic of Abkhazia, Viache On April 17, 2012 in Moscow, the Minister of Foreign Affairs of the At the meeting a wide range of issues of mutual interest were disc
MFAs	Abkhazia	2012-04-18	3	On April 18- 2012 a press conference of the	en	http://mfaapsny.org/en/information/?ID=53	18.04.2012 On April 18, 2012 a press conference of the Minister of Foreign Aff On April 18, 2012 a press conference of the Minister of Foreign Aff Journalists were interested in a number of issues, in particular, the The head of the Abkhazian Foreign Ministry, told the journalists ab According to him, the Geneva discussions are the only place whe Journalists at the press conference also touched the issue of Geo The visit of the delegation of the Ministry of Foreign Affairs of Abkh The press conference lasted more than an hour.
							23.04.2012

- Spreadsheet
- Folder of txt files
- All articles within a single txt file
- Document-term matrix
- ...

2012-09-20 - AbkhaziaMfa - 88 - The Ministry of Foreign Affairs of the Republic of Abkhazia sends a congratulato.txt
2012-09-21 - AbkhaziaMfa - 89 - Information by the Foreign Ministry of the Republic of Abkhazia about the Proces.txt
2012-09-24 - AbkhaziaMfa - 90 - Viacheslav Chirikba met with the UNDP Resident Representative Jamie McGoldrick.txt
2012-09-25 - AbkhaziaMfa - 91 - Deputy Foreign Minister Irakli Khintba interviewed by the geopolitical magazine .txt
2012-09-26 - AbkhaziaMfa - 92 - Meeting of the Deputy Ministers of Foreign Affairs of the Republic of Abkhazia I.txt
2012-09-27 - AbkhaziaMfa - 93 - Acting Minister, Deputy Foreign Minister of the Republic of Abkhazia Irakli Khin.txt
2012-09-30 - AbkhaziaMfa - 95 - A diplomatic reception on the occasion of the Victory and Independence Day of th.txt
2012-10-02 - AbkhaziaMfa - 96 - The Ministry of Foreign Affairs of the Republic of Abkhazia sent a congratulator.txt
2012-10-02 - AbkhaziaMfa - 97 - Acting Minister, Deputy Minister of Foreign Affairs of the Republic of Abkhazia .txt
2012-10-04 - AbkhaziaMfa - 98 - The Ministry of Foreign Affairs of the Republic of Abkhazia sent a congratulator.txt
2012-10-08 - AbkhaziaMfa - 99 - Interview of the Deputy Foreign Minister of the Republic of Abkhazia Irakly Khin.txt
2012-10-11 - AbkhaziaMfa - 100 - On October 11, 2012 there will be the XXI round of Geneva discussions on securi.txt

Why bother?

- "identify widespread patterns of naturally occurring language and rare but telling examples, both of which may be overlooked by a small-scale analysis."

Baker, Paul, and Tony McEnery. 2005. "A Corpus-Based Approach to Discourses of Refugees and Asylum Seekers in UN and Newspaper Texts." *Journal of Language & Politics* 4 (2): 197–226.

- Find the needle in the haystack
- Characterise the haystack

Hopkins, Daniel J., and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–47. doi:10.1111/j.1540-5907.2009.00428.x.

Why bother?

- By treating the internet as an inordinate mass of contents that can be superficially explored through search engines and serendipitous findings, we are missing out:
  - we cannot see trends, we may miss information
- Methodological rigour, replicability, and such

Some examples

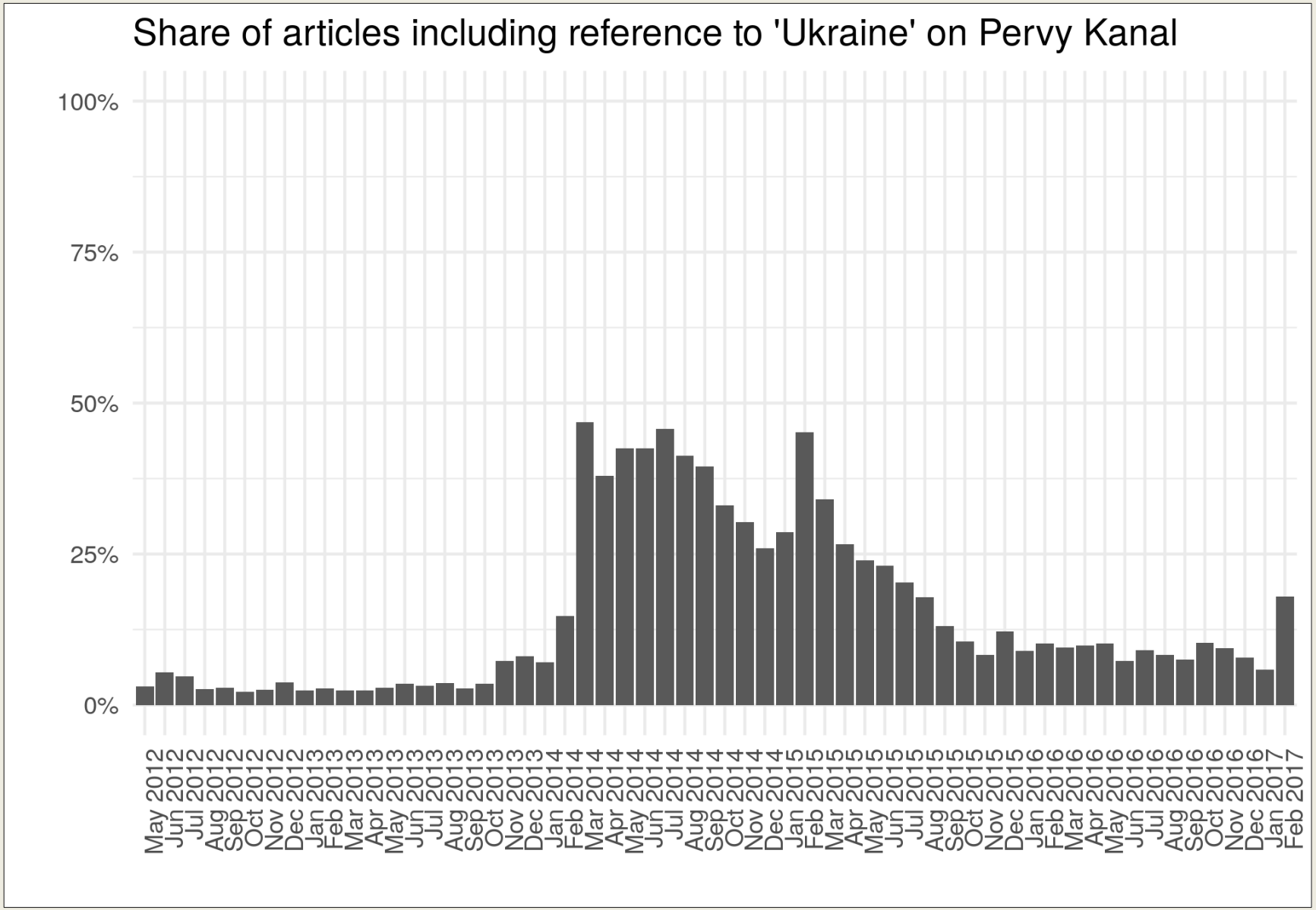
(the point is not “look at what I have done”, but rather “could something like this be useful to answer your own research questions”)

Examples based on word frequency

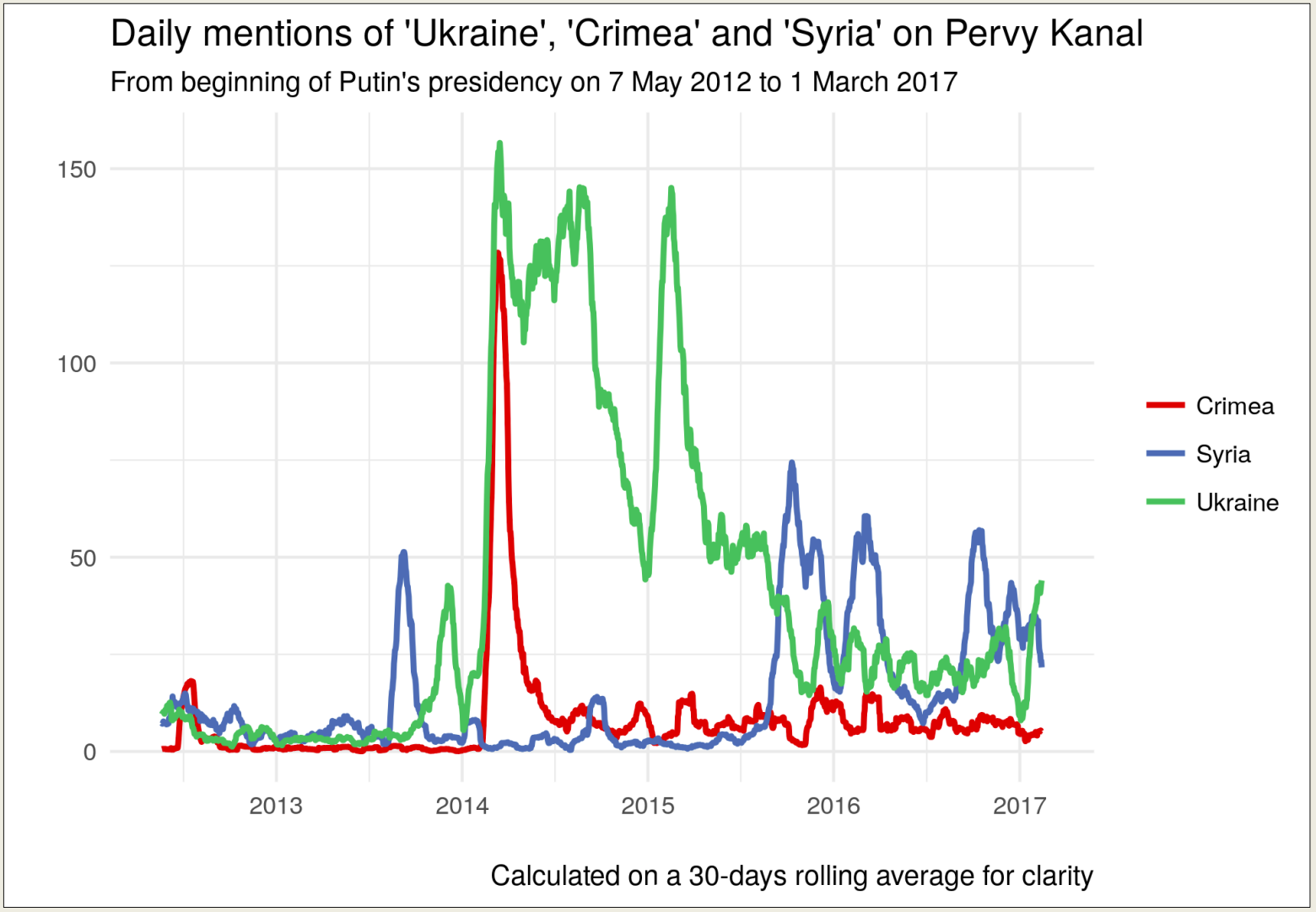
"The most valuable use of studies of content [...] is in noting trends and changes in content"
(Albig, 1938, p. 349)

Albig, William. 1938. "The Content of Radio Programs, 1925-1935." Social Forces 16 (3): 338–49. doi:10.2307/2570805.

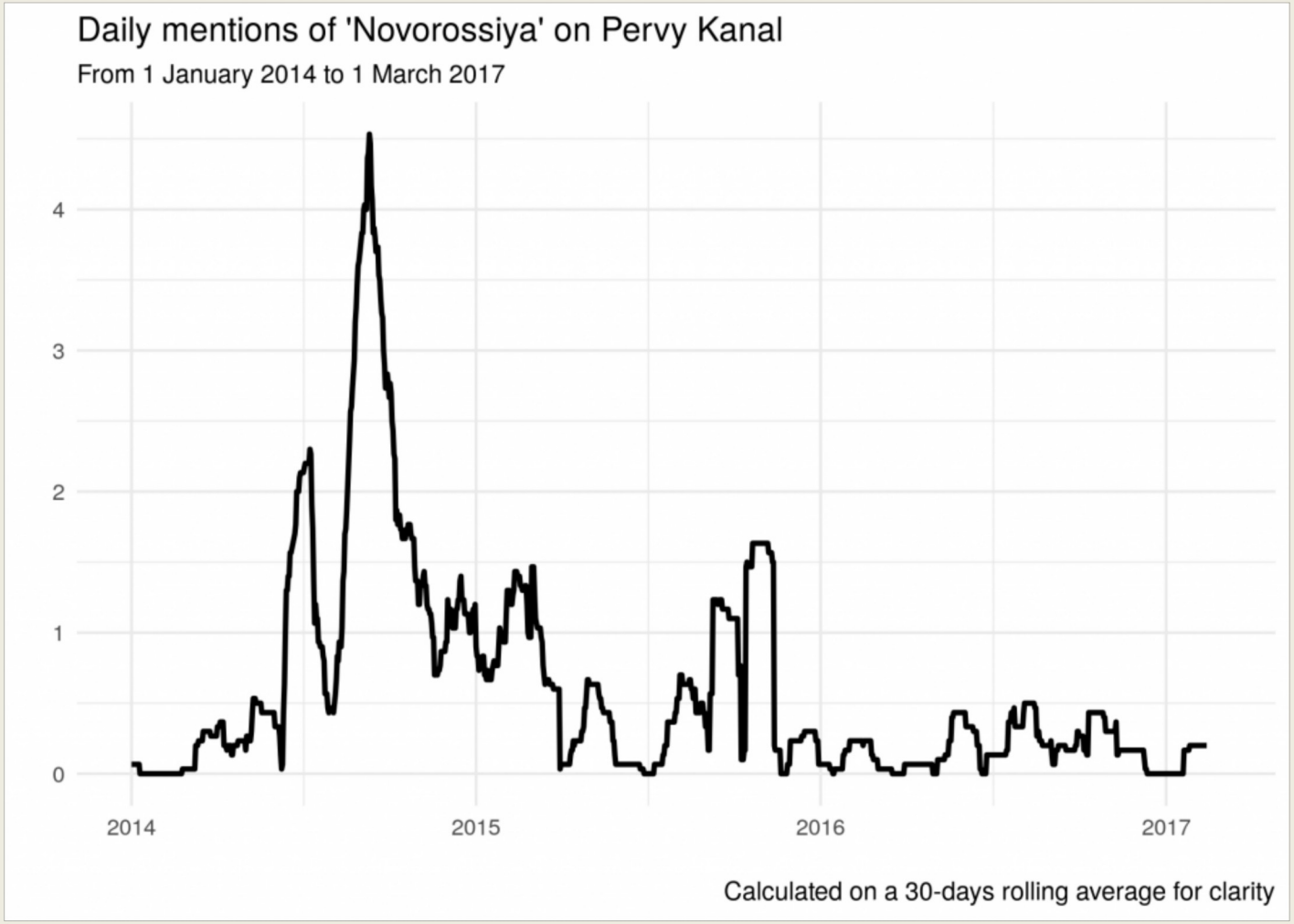
Example #1: Russian media and Ukraine



Example #2: Russian media Ukraine/Crimea/Syria

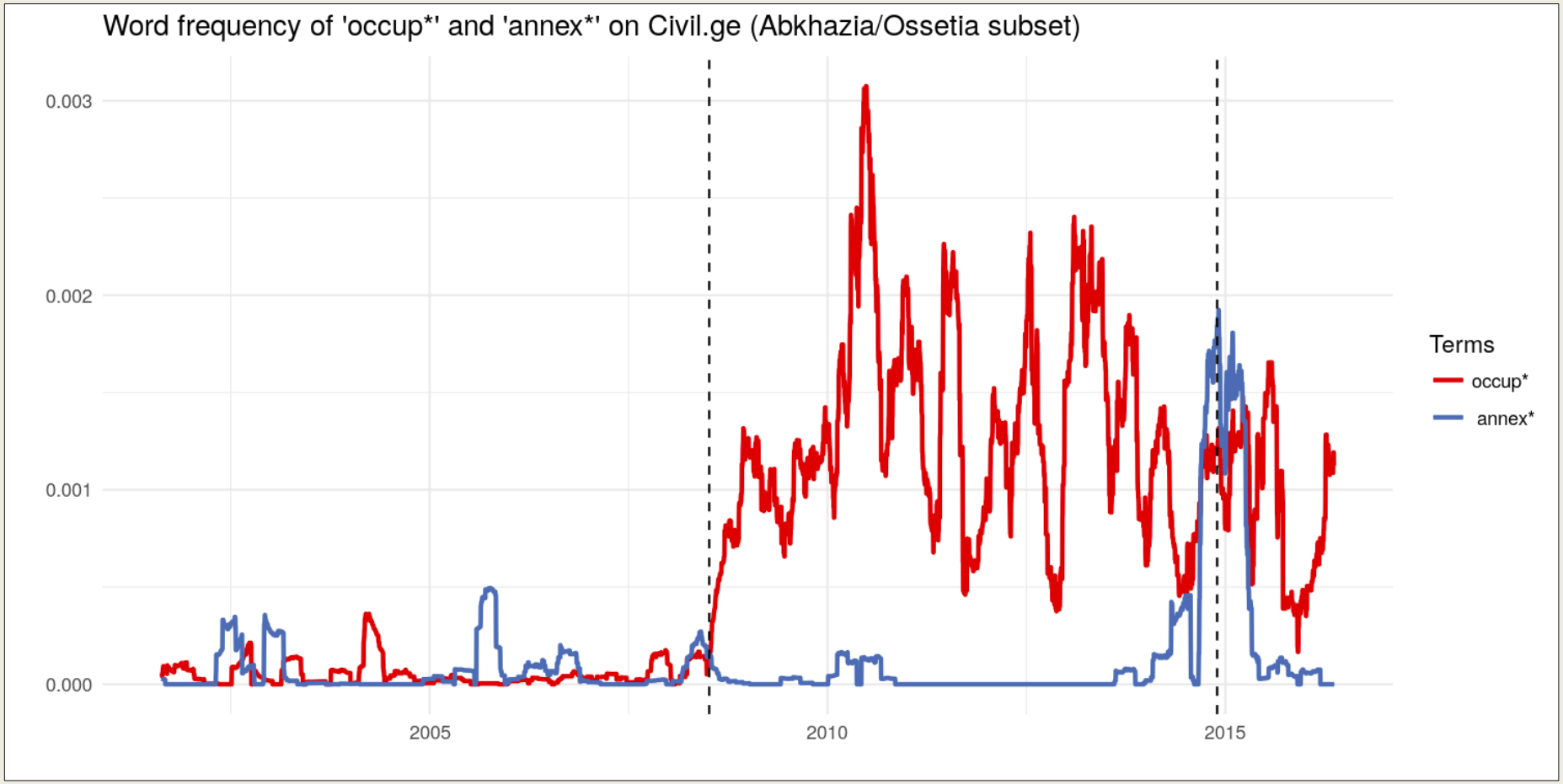


Example #3: Novorossiya

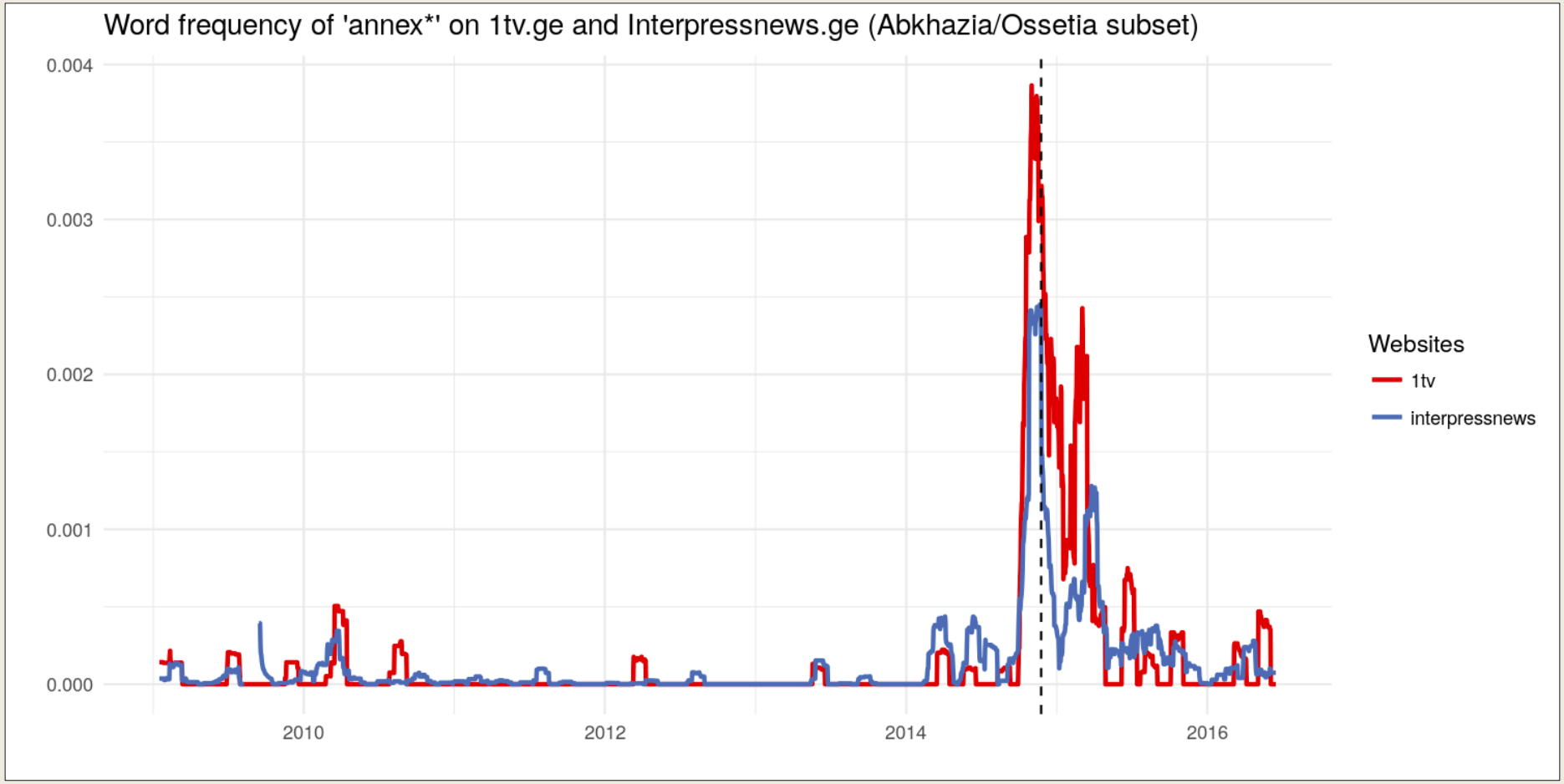


http://www.giorgiocomai.eu/2017/03/20/word-frequency-of-ukraine-crimea-dnrlnr-and-novorossiya-on-1tv-ru/

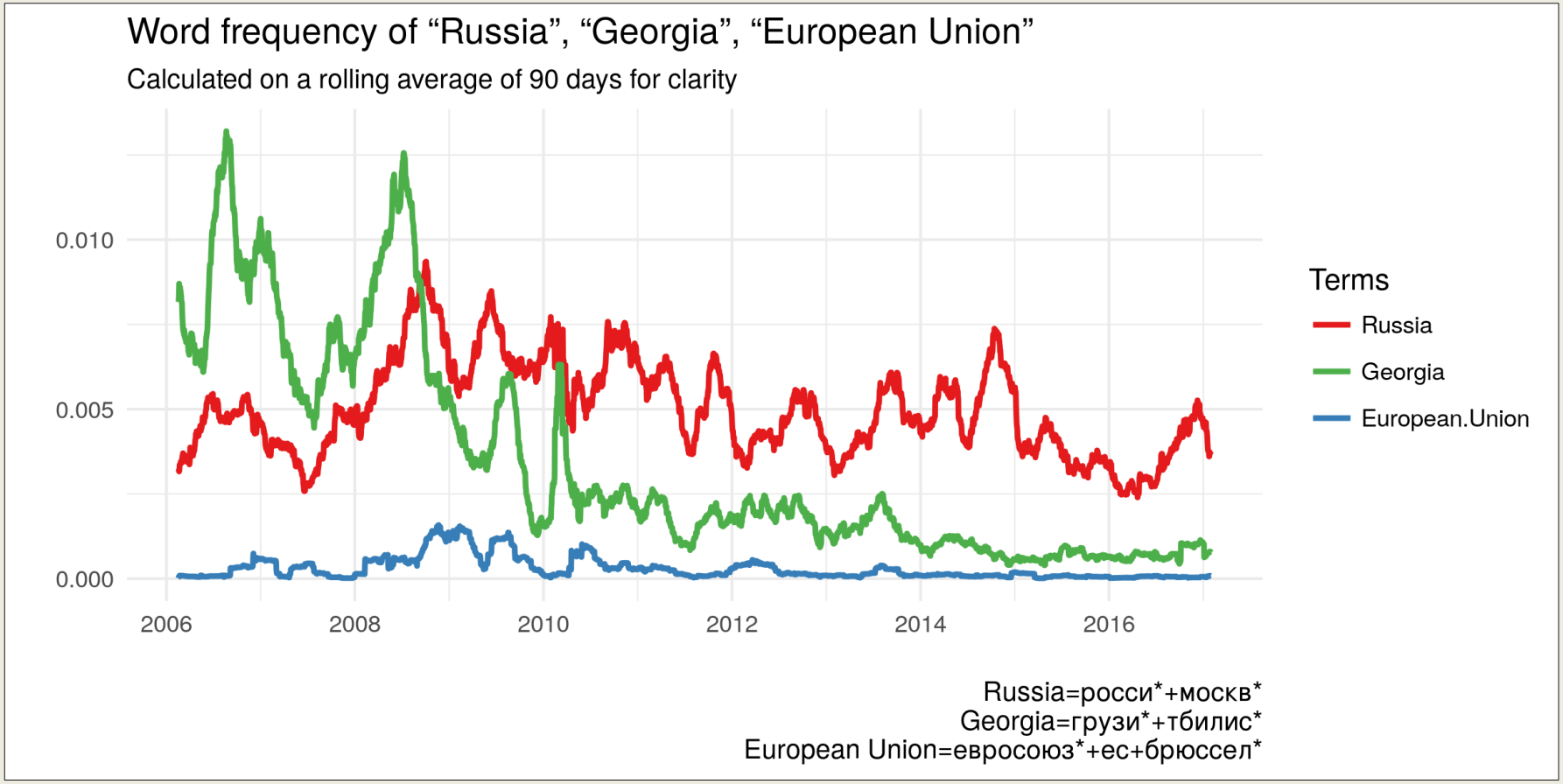
Local media


Word frequency of 'occup*' and 'annex*' on Civil.ge (Abkhazia/Ossetia subset)
0.003
0.002
0.001
0.000
2005
2010
2015
Terms
occup*
annex*

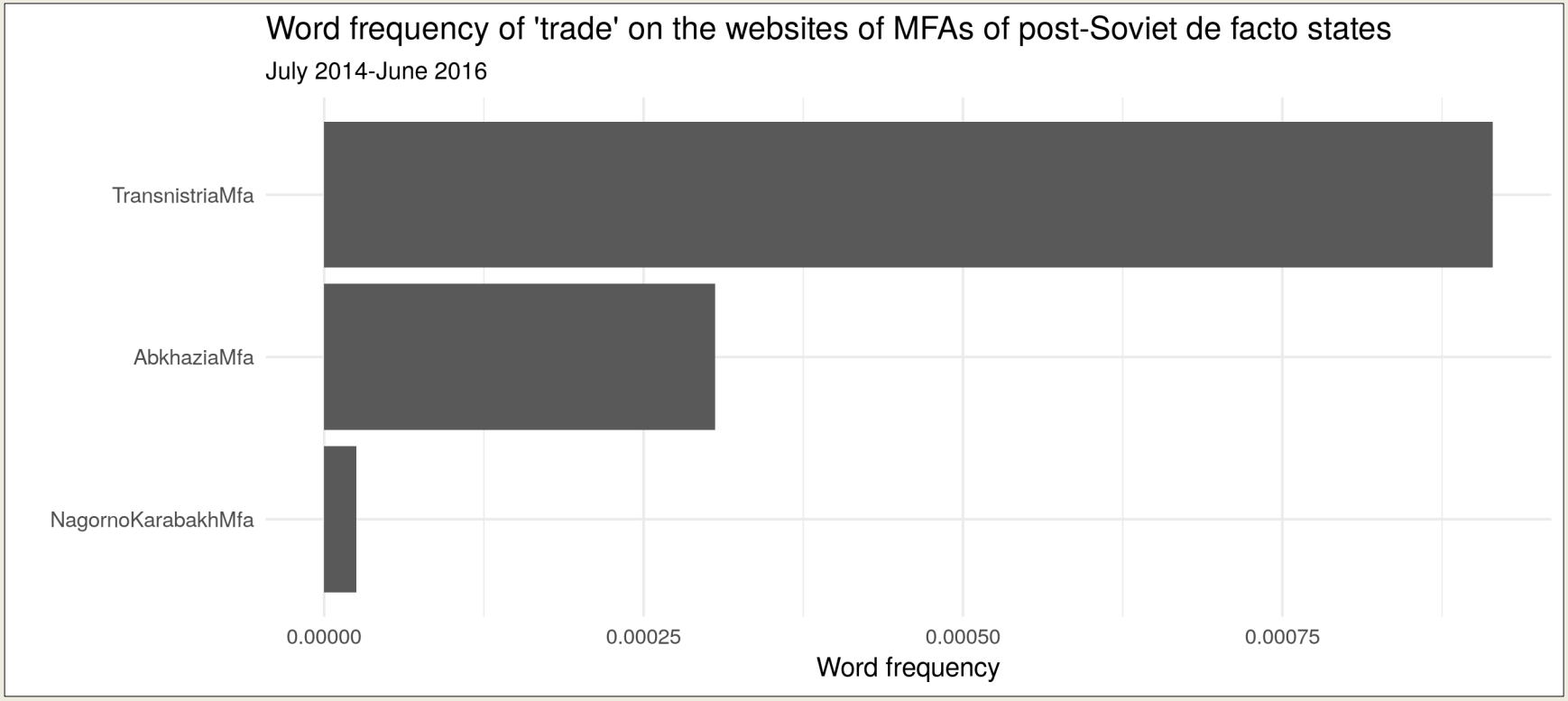
Not necessarily in ‘mainstream’ languages



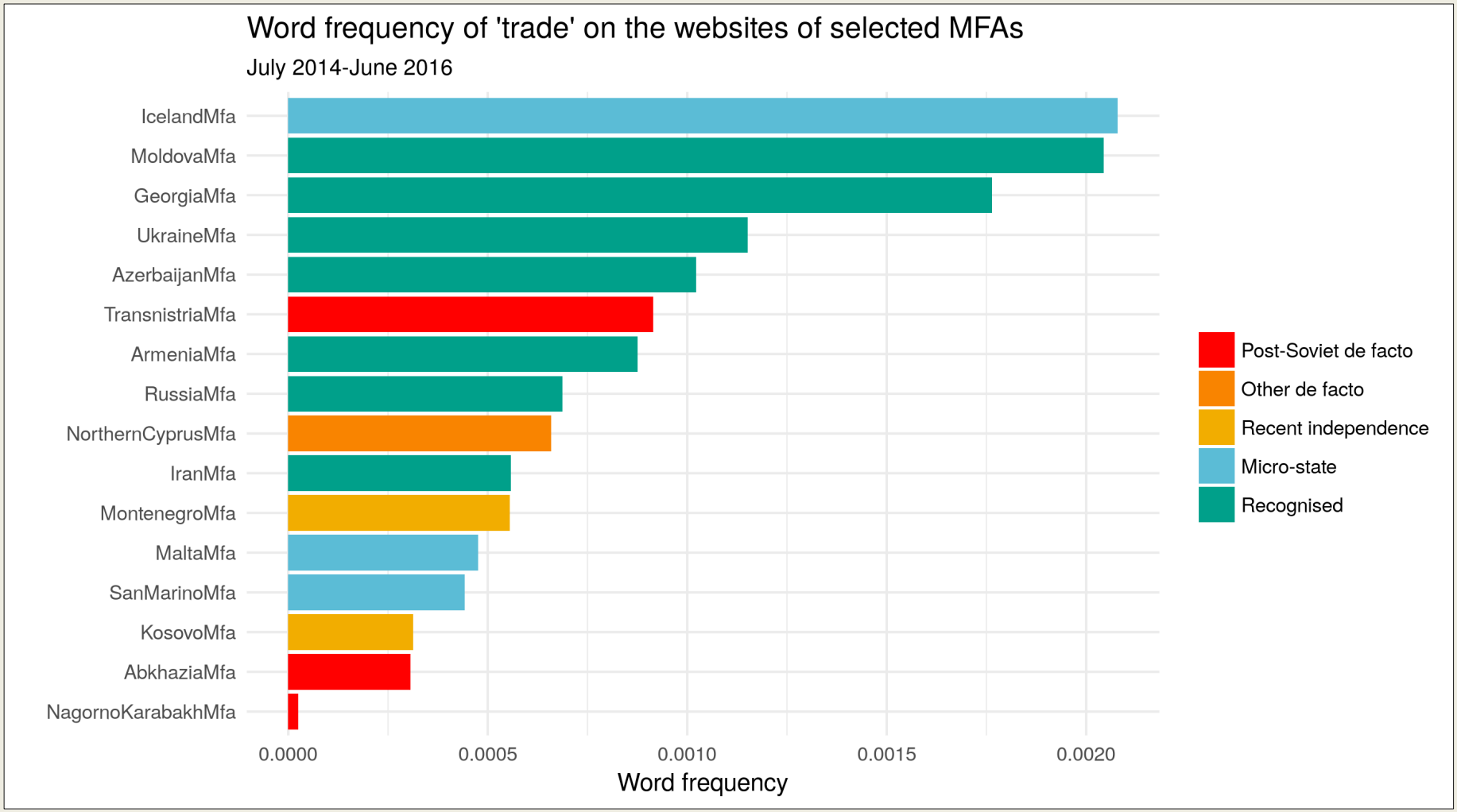
Or from sources that may not appear in established databases



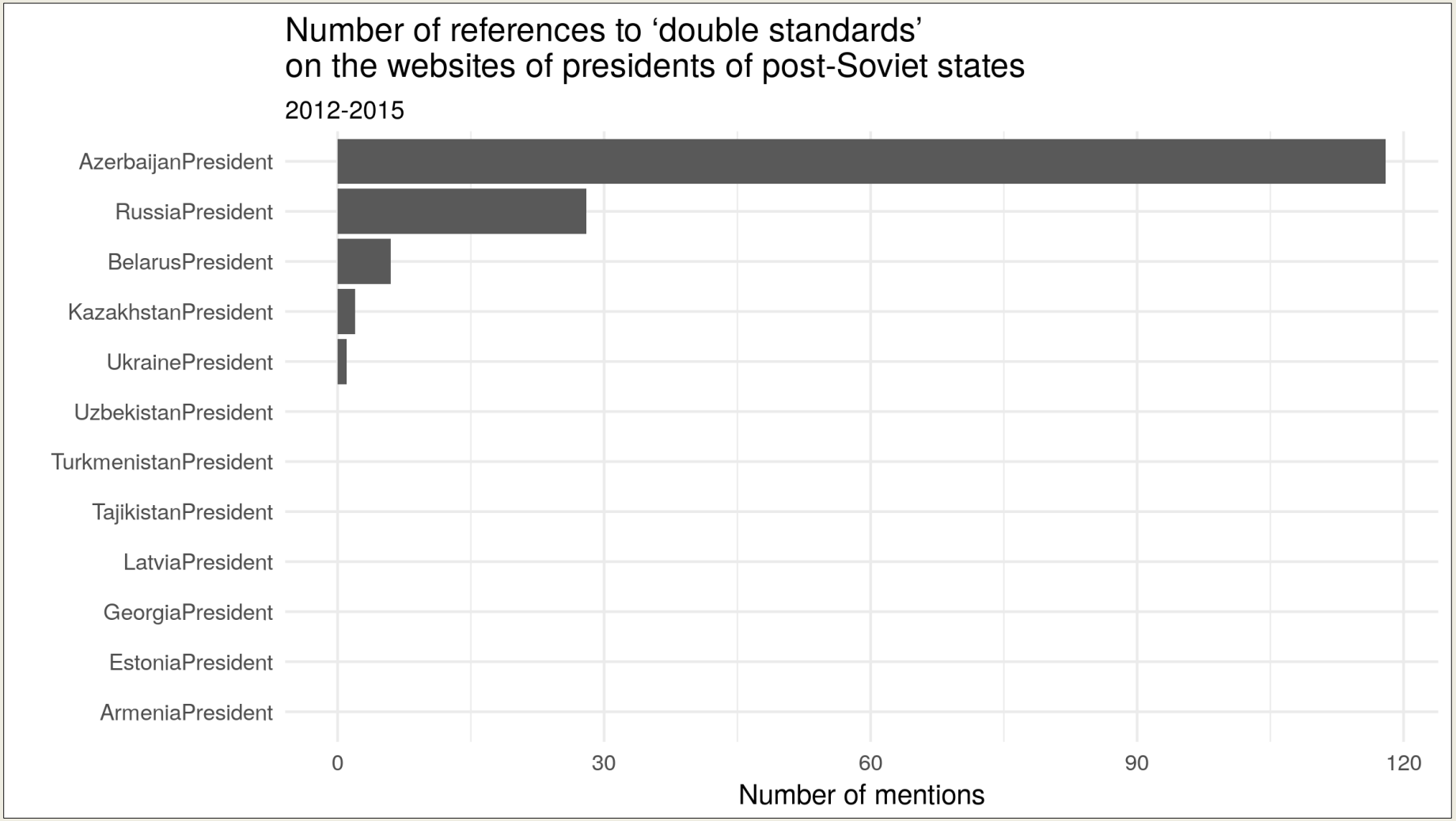
Such as institutions in de facto states (or local institutions anywhere, really)



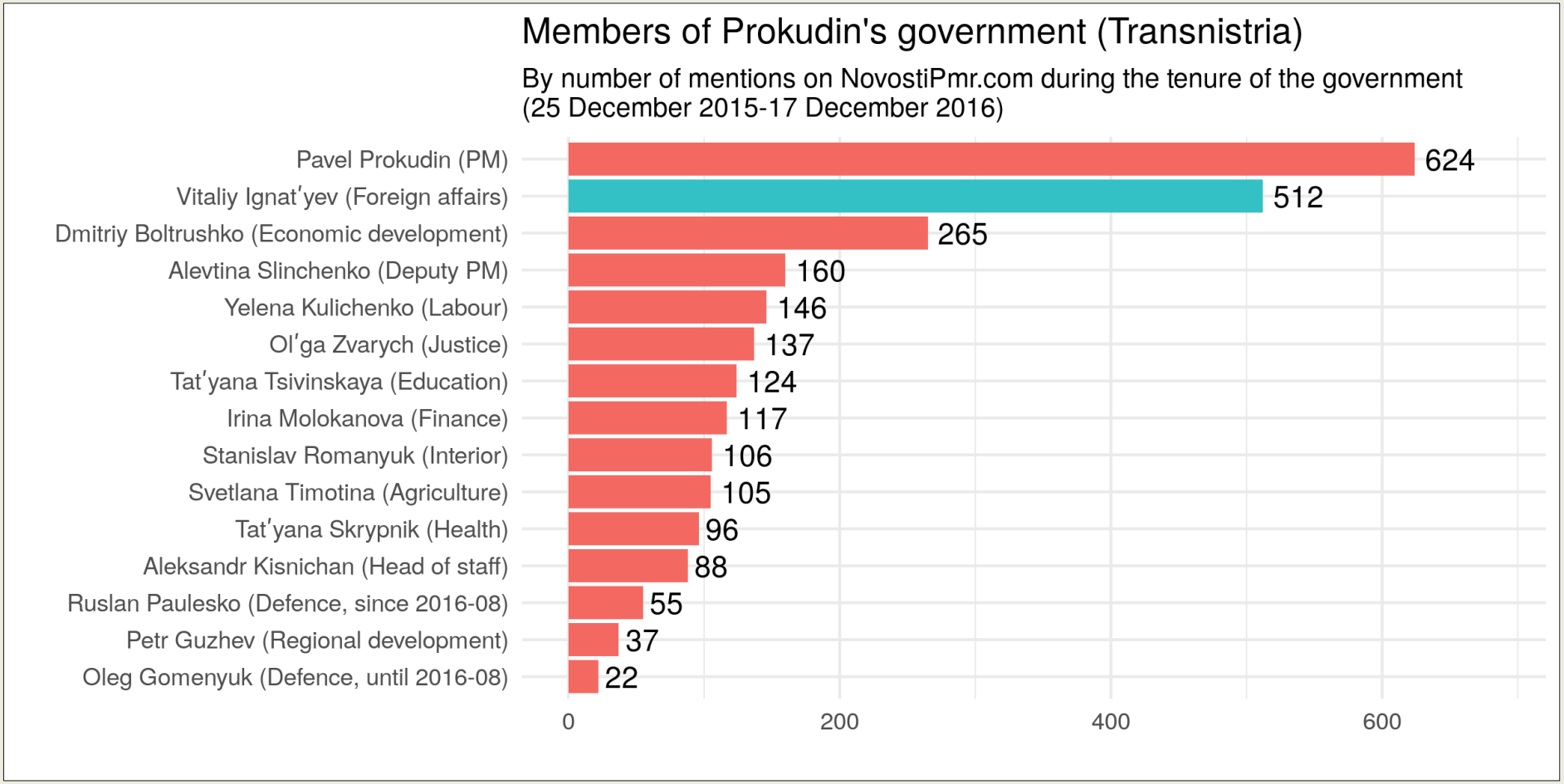
More cases


Word frequency of 'trade' on the websites of selected MFAs
July 2014-June 2016
IcelandMfa
MoldovaMfa
GeorgiaMfa
UkraineMfa
AzerbaijanMfa
TransnistriaMfa
ArmeniaMfa
RussiaMfa
NorthernCyprusMfa
IranMfa
MontenegroMfa
MaltaMfa
SanMarinoMfa
KosovoMfa
AbkhaziaMfa
NagornoKarabakhMfa
0.0000
0.0005
0.0010
0.0015
0.0020
Word frequency
Post-Soviet de facto
Other de facto
Recent independence
Micro-state
Recognised

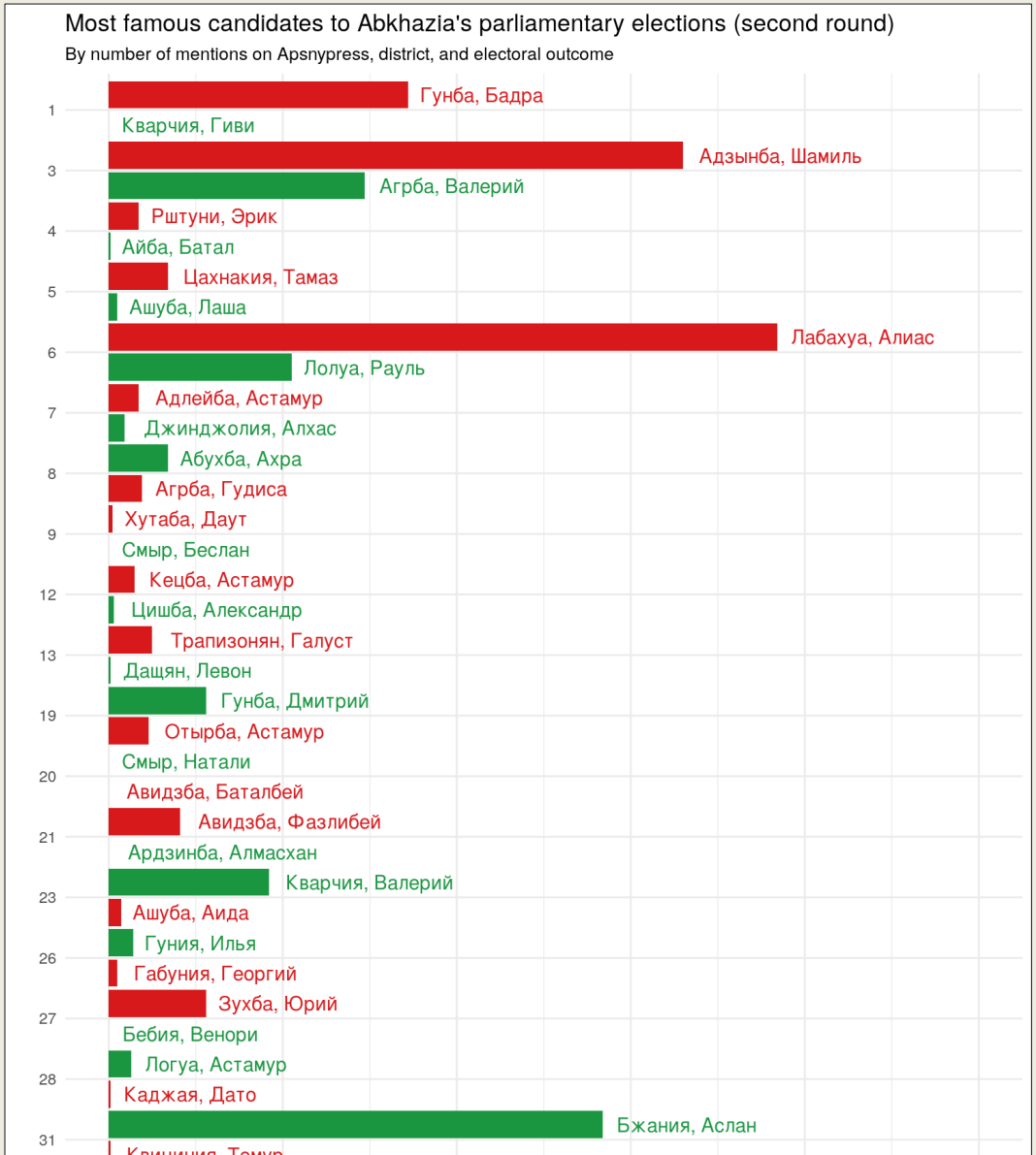
To present case selection



To point at visibility of politicians, or of the institution they represent



Or visibility of candidates



http://www.giorgiocomai.eu/2017/03/31/abkhazias-parliamentary-elections-not-for-the-famous/

Or simply to subset materials and analyse them qualitatively, find data, etc.

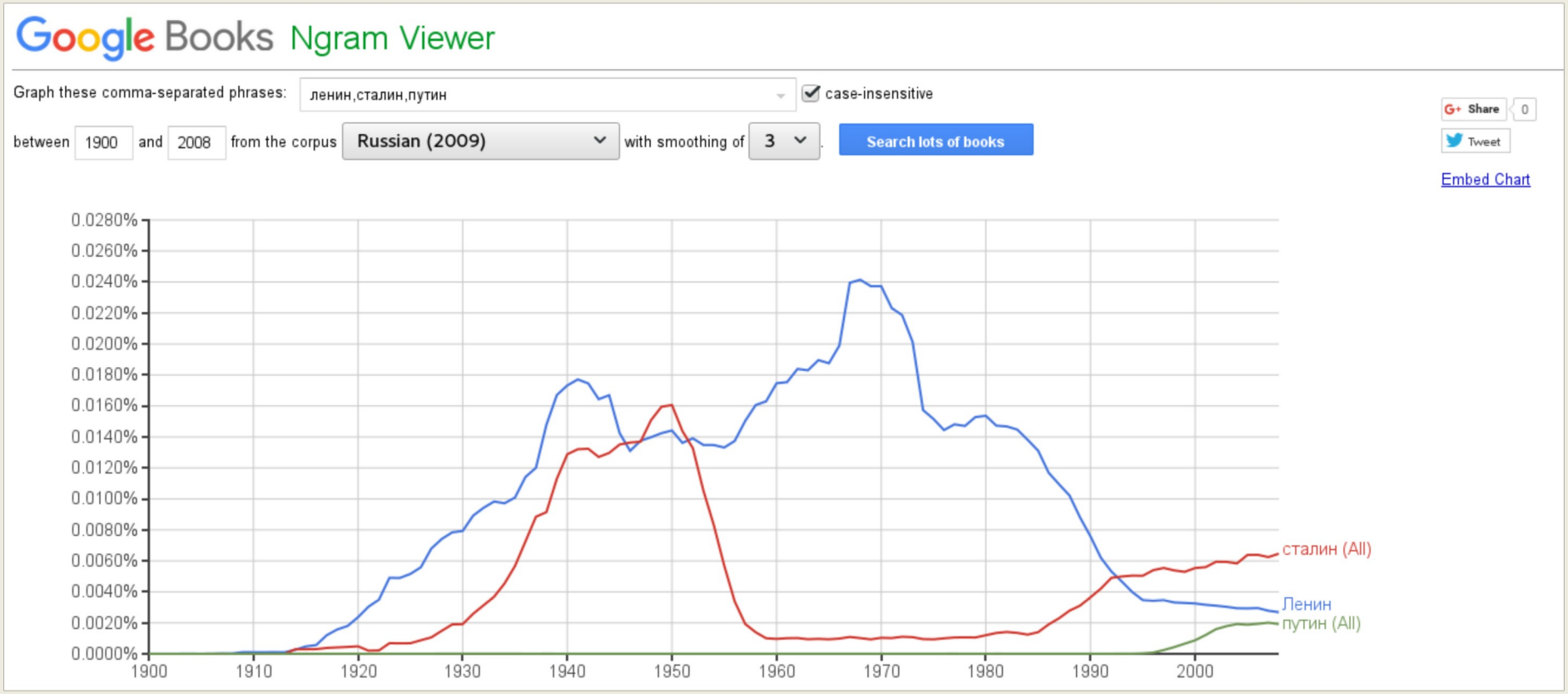
Why is this approach so uncommon in area studies?

- (area studies) researchers don't know how to go about it
- It is technically complicated and time-consuming
- Epistemological issues (?)
- ?

Decades after the technique was established, and in spite of technological advancements "content analysis is still an expensive research tool. [...] And it is so even in computer-assisted content analysis. [...] Computer-aided content analysis is still time consuming." (Franzosi 2008, p. XXXV)

Franzosi, Roberto, ed. 2008. *Content Analysis*. Sage Benchmarks in Social Research Methods. London: SAGE.

Are we just lazy?



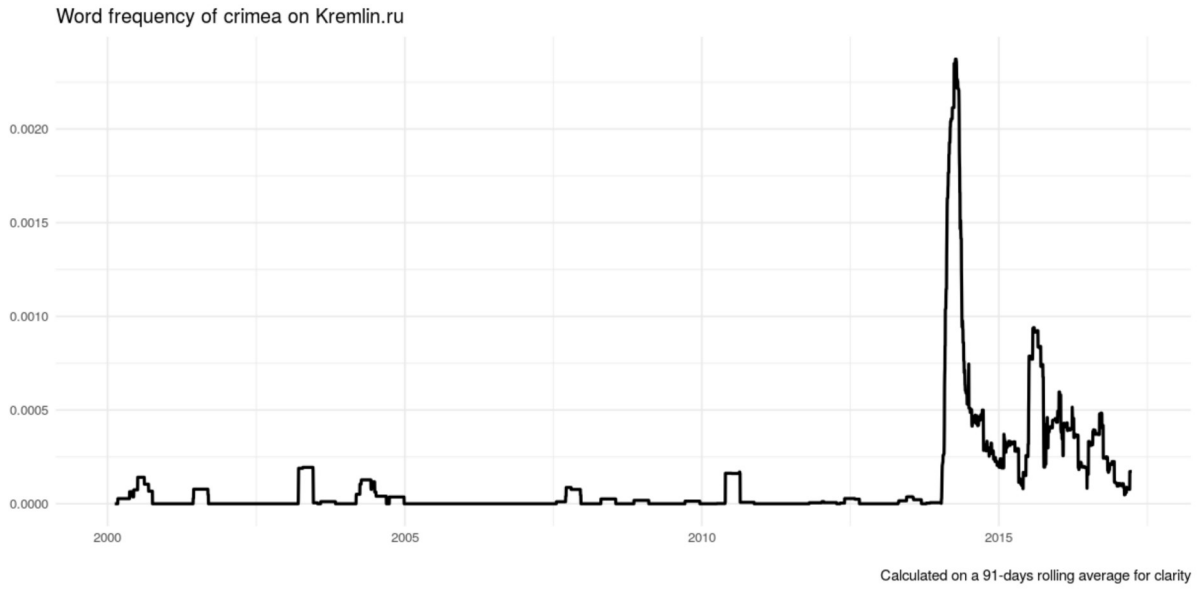
If it was easier, would more area-study researchers use this approach?

If it was easier, would more area-study researchers use this approach?

Word frequency on Kremlin.ru

Term to be analysed

crimea



https://giocomai.shinyapps.io/kremlin/

If it was easier, would more area-study researchers use this approach?

Official websites of South Caucasus presidents

Select type of graph

- (•) Time series
- ( ) Barchart

Terms to be analysed*

europ*

startDate

2014-06-01

endDate

2016-05-30

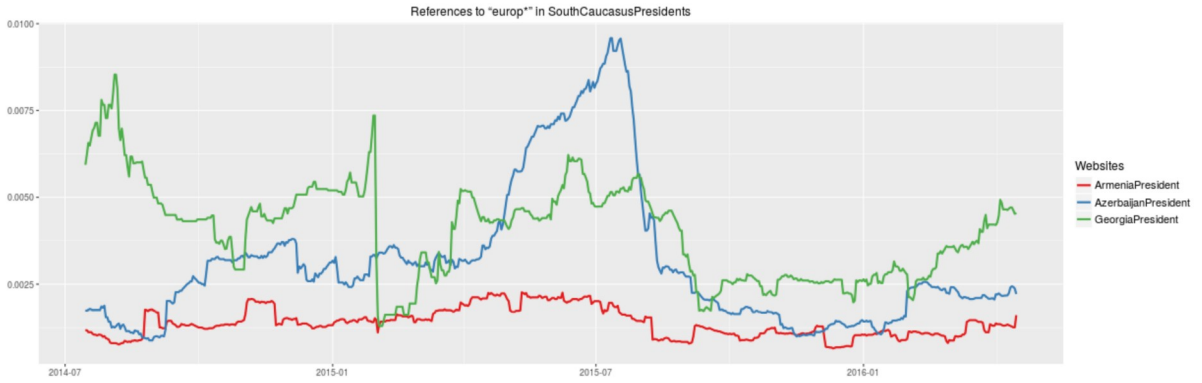
Rolling average (days)

1 — 90

1 10 19 28 37 46 55 64 73 82 90

- [ ] Smooth
- [ ] Linear smooth
- [ ] Show interactive time series
- [x] Show keywords in context

* If more than one, terms must be comma-separated; to merge mutliple terms in calculations, use the format: Russia=Russia*+Moscow+Putin, Europe=Europ*+EU+Brussels

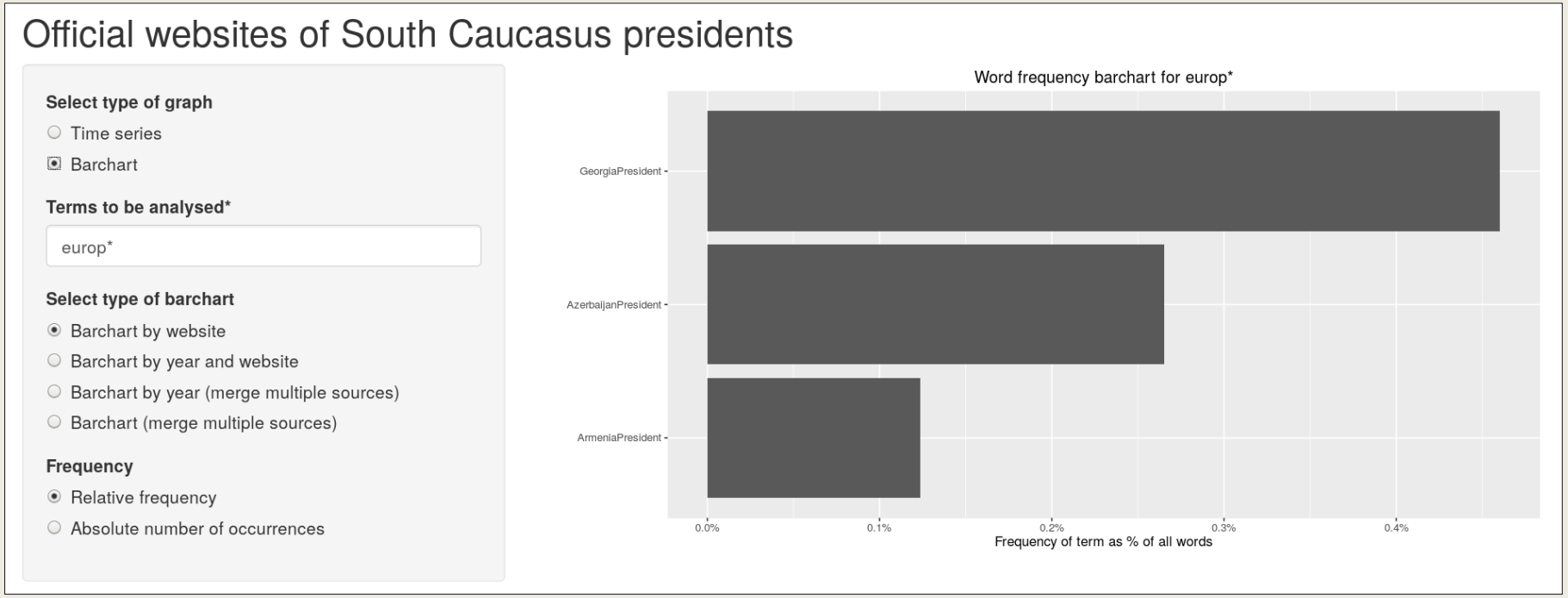


Show 25 entries

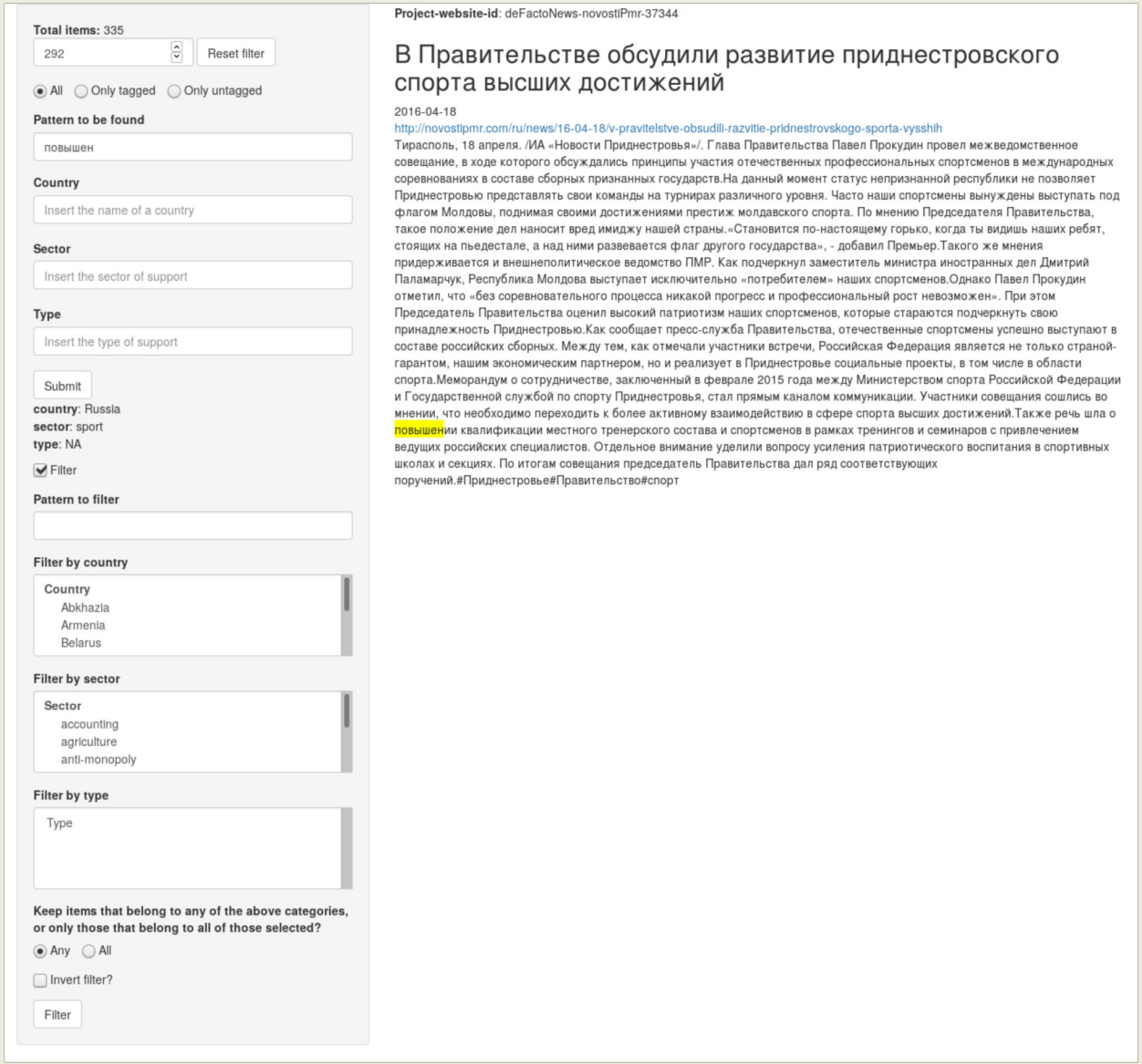
Search:

Date	contextPre	keyword	contextPost	Source
2014-06-02	consolidating the collaboration between Armenia and the	European	Union, as well as at further	President Serzh Sargsyan held meeting with leadership of diplomatic service
2014-06-02	reinforcing and advancing the bilateral relations with	European	countries. The President of Armenia drew	President Serzh Sargsyan held meeting with leadership of diplomatic service
2014-06-05	figure skating, the multiple world and	European	champion, the Russian public	President took part in groundbreaking ceremony of new figure skating school and visited companies of Grand Holding

If it was easier, would more area-study researchers use this approach?



Maybe also to share qualitative coding, notes, etc.?



Total items: 335

292 Reset filter

All Only tagged Only untagged

Pattern to be found

повышен

Country

Insert the name of a country

Sector

Insert the sector of support

Type

Insert the type of support

Submit

country: Russia
sector: sport
type: NA

Filter

Pattern to filter

Filter by country

Country
Abkhazia
Armenia
Belarus

Filter by sector

Sector
accounting
agriculture
anti-monopoly

Filter by type

Type

Keep items that belong to any of the above categories, or only those that belong to all of those selected?

Any All

Invert filter?

Filter

Project-website-id: deFactoNews-novostiPmr-37344

В Правительстве обсудили развитие приднестровского спорта высших достижений

2016-04-18
http://novostipmr.com/ru/news/16-04-18/v-pravitelstve-obsudili-razvitie-pridnestrovskogo-sporta-vysshih
Тирасполь, 18 апреля. /ИА «Новости Приднестровья»/. Глава Правительства Павел Прокудин провел межведомственное совещание, в ходе которого обсуждались принципы участия отечественных профессиональных спортсменов в международных соревнованиях в составе сборных признанных государств.На данный момент статус непризнанной республики не позволяет Приднестровью представлять свои команды на турнирах различного уровня. Часто наши спортсмены вынуждены выступать под флагом Молдовы, поднимая своими достижениями престиж молдавского спорта. По мнению Председателя Правительства, такое положение дел наносит вред имиджу нашей страны.«Становится по-настоящему горько, когда ты видишь наших ребят, стоящих на пьедестале, а над ними развевается флаг другого государства», - добавил Премьер.Такого же мнения придерживается и внешнеполитическое ведомство ПМР. Как подчеркнул заместитель министра иностранных дел Дмитрий Паламарчук, Республика Молдова выступает исключительно «потребителем» наших спортсменов.Однако Павел Прокудин отметил, что «без соревновательного процесса никакой прогресс и профессиональный рост невозможен». При этом Председатель Правительства оценил высокий патриотизм наших спортсменов, которые стараются подчеркнуть свою принадлежность Приднестровью.Как сообщает пресс-служба Правительства, отечественные спортсмены успешно выступают в составе российских сборных. Между тем, как отмечали участники встречи, Российская Федерация является не только страной-гарантом, нашим экономическим партнером, но и реализует в Приднестровье социальные проекты, в том числе в области спорта.Меморандум о сотрудничестве, заключенный в феврале 2015 года между Министерством спорта Российской Федерации и Государственной службой по спорту Приднестровья, стал прямым каналом коммуникации. Участники совещания сошлись во мнении, что необходимо переходить к более активному взаимодействию в сфере спорта высших достижений.Также речь шла о повышении квалификации местного тренерского состава и спортсменов в рамках тренингов и семинаров с привлечением ведущих российских специалистов. Отдельное внимание уделили вопросу усиления патриотического воспитания в спортивных школах и секциях. По итогам совещания председатель Правительства дал ряд соответствующих поручений.#Приднестровье#Правительство#спорт

Concluding remarks

- Actually giving the chance to explore and conduct analysis of the datasets we created (both quantitative and qualitative) opens new possibilities for feedback and alternative explanations
- Creating textual datasets based on websites (media, institutions, etc.) can be easy
- And if we all share them, then it's even easier, and we can build upon each other's work
- If it was easy enough, would researchers do it?

Text-mining the post-Soviet web

Giorgio Comai

Dublin City University

@giocomai

Comai, Giorgio (forthcoming, 2017). "Quantitative Analysis of Web Contents in Support of Qualitative Research. Examples from the Study of Post-Soviet de Facto States." *Studies of Transition States and Societies.*

http://giorgiocomai.eu/

https://github.com/giocomai/castarter
[code is being rewritten, not fully functional]
https://giocomai.shinyapps.io/kremlin/
https://giocomai.shinyapps.io/kremlinregex/
https://giocomai.shinyapps.io/SouthCaucasusPresidents/